LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models

Year: 2023 | Citations: 304 | Authors: Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang

Abstract

This work aims to learn a high-quality text-to-video (T2V) generative model by leveraging a pre-trained text-to-image (T2I) model as a basis. It is a highly desirable yet challenging task to simultaneously (a) accomplish the synthesis of visually realistic and temporally coherent videos while (b) preserving the strong creative generation nature of the pre-trained T2I model. To this end, we propose LaVie, an integrated video generation framework that operates on cascaded video latent diffusion models, comprising a base T2V model, a temporal interpolation model, and a video super-resolution model. Our key insights are two-fold: (1) We reveal that the incorporation of simple temporal self-attentions, coupled with rotary positional encoding, adequately captures the temporal correlations inherent in video data. (2) Additionally, we validate that the process of joint image-video fine-tuning plays a pivotal role in producing high-quality and creative outcomes. To enhance the performance of LaVie, we contribute a comprehensive and diverse video dataset named Vimeo25M, consisting of 25 million text-video pairs that prioritize quality, diversity, and aesthetic appeal. Extensive experiments demonstrate that LaVie achieves state-of-the-art performance both quantitatively and qualitatively. Furthermore, we showcase the versatility of pre-trained LaVie models in various long video generation and personalized video synthesis applications. Project page: https://github.com/Vchitect/LaVie/.